

Data-driven Classification of Linguistic Styles in Spoken Dialogues

Thomas Portele

Philips Research Laboratories Aachen

Thomas.Portele@philips.com

Abstract

Language users have individual linguistic styles. A spoken dialogue system may benefit from adapting to the linguistic style of a user in input analysis and output generation. To investigate the possibility to automatically classify speakers according to their linguistic style three corpora of spoken dialogues were analyzed. Several numerical parameters were computed for every speaker. These parameters were reduced to linguistically interpretable components by means of a principal component analysis. Classes were established from these components by cluster analysis. Unseen input was classified by trained neural networks with varying error rates depending on corpus type. A first investigation in using special language models for speaker classes was carried out.

1 Motivation

Within spoken dialogues the participants make individual use of the linguistics of the pertinent language. On one hand, each participant has a linguistic style as an important element of his/her personality (Pieper, 1979; Walker et al., 1997). The quantitative analysis of linguistic style (counting and comparing) has been used in linguistics and literature for a long time to determine authorship of written texts (Mendenhall, 1887). Lehman and Carbonell (1989) describe a system for written natural language queries that tries to adapt to the user's grammar by starting from a simple basic grammar and relaxing and augmenting it if a user provides uninterpretable input. They found significant differences among the active linguistic patterns generated by different users, but each user was fairly consistent across sessions spanning several days. For spoken dialogue systems, this aspect of style can be important to optimize the analysis of a user's *input*.

On the other hand, social bonding is performed by adapting to a common interaction style (Brown and Levinson, 1987; Okada et al., 1999). It has been shown that stylistic elements of one participant are adopted by other participants (Fais and Loken-kim, 1995; Gustafson et al., 1997). Studies have indicated that variations between conversations is high but low within conversations (Brennan, 1996) because people mark their shared conceptualizations by using the same term, *lexical entrain-*

ment. A spoken dialogue system can use stylistic information to adapt its *output* behavior.

Further determiners of style are the domain or genre (Karlgrén and Cutting, 1994; Wolters and Kirsten, 1999), the modalities (speech only vs. speech and visual interfaces) (Fais et al., 1996; Oviatt et al., 1994), and the degree of interactivity (Oviatt and Cohen, 1991) which are more or less determined by the application scenario of a spoken dialogue system.

One specific aspect of spoken input is its larger irregularity. For written texts, robust parsers can be employed to obtain style-relevant information (Karlgrén, 1994; Paiva, 2000), while for spontaneous speech simpler measurements have to be used like part-of-speech tags (Ries, 1999).

Klarner (1997) investigated stylistic differences of speakers in the Verbmobil dialogue corpus in order to improve speech recognition by using speaker-type dependent language models. The achieved reduction in perplexity, however, is relatively low.

For the research project *SmartKom* funded by the German ministry of research (BMBF) (Wahlster et al., 2001) a module is being developed that constructs and maintains a model of human-computer interaction. One part models the interaction style of the user (experience with the system, experience with the task, preferred modalities for input and output), the other part the linguistic style. Both parts are supposed to make use of stereotypes (Rich, 1979).

The experiments described below explore the possibility to consistently extract linguistic parameters from spoken dialogues, to use these parameters in order to group speakers into several classes, and to train learning algorithms that classify users by their parameter values.

2 Corpora

The task of a spoken dialogue system is to engage in spoken human-computer interaction. It is well known that spoken human-computer interaction differs from its human-human counterpart in various dimensions (Doran et al., 2001) including linguistic complexity. For the purpose of this investigation three sources were exploited: a corpus of task-dependent human-human interactions (negotiation dialogues), a corpus of free human-human

conversations, and a corpus of human–computer interactions. For all corpora the part–of–speech information for each word was automatically annotated by the IMS tree tagger (Schmid, 1994) using the STTS tagset (Schiller et al., 1995).

Verbmobil The Verbmobil (VM) corpus (Wahlster, 1993) is one of the largest spoken dialogue corpora available for German. It contains spontaneous speech human–human dialogues in the appointment negotiation and travel planning domain. The corpus used for this investigation has data from 837 speakers (24569 turns with 448737 words, av. 29.35 turns per speaker).

CallHome The CallHome (CH) corpus (Linguistic Data Consortium, 1997) contains 80 dialogues of 10 minutes unconstrained conversation between two humans over the telephone. The corpus has utterances from 160 speakers (17744 turns with 145552 words, av. 110.9 turns per speaker).

TABA The TABA corpus contains human–computer dialogues in the domain of train timetable information (Aust et al., 1995). The transcription was done automatically by the speech recognizer of the dialogue system. As the recognizer can only recognize words present in the pertinent recognition lexicon and may be subject to errors it is likely that the corpus sometimes does not contain the actual words uttered by the speaker contrary to the other corpora. The corpus consists of 5200 dialogues (33568 turns with 90377 words, av. 6.45 turns per speaker).

3 Method and Results

3.1 Parameter values

For each turn a set of parameter values was computed. The STTS tagset consists of more than 50 different tags. In order to obtain reasonable results the STTS tagset was collapsed to a set with 12 classes. Their frequency distributions (henceforth Cxxx, e.g. CART for the frequency of articles) indicate the differences between the corpora (Figure 1). While the TABA corpus mainly contains nouns, prepositions, and particles, the two human–human corpora have many pronouns (pronominal referencing is possible due to longer contexts), verbs (varying tasks need task names, sentences in longer utterances are less likely to be elliptic), and adverbs (an utterance is put in relation to its context). The TABA corpus features nearly no interjections, while the number of numerals in the CH corpus is rather low (in the other corpora times and dates were explicit elements of the tasks). The length distributions (Lxxx) are similar with the exception of numerals in the VM corpus (dates are quite long in German, e.g. “zweiundzwanzigster” (22nd)). An additional set of parameters is the relative frequency of the different classes in phrase–final position (Fxxx) (Klarner, 1997).

Tag	Meaning	VM	CH	TABA
ADJ	adjectives	4.7	5.00	2.2
ADV	adverbs	15.9	17.4	6.6
PRP	prepositions	9.5	4.9	22.3
ART	articles	5.5	5.6	1.2
NUM	numerals	5.3	0.9	7.3
ITJ	interjections	3.4	2.0	0.1
KON	conjunctions	4.0	6.9	1.1
NOM	nouns	15.4	12.2	31.8
PRO	pronouns	17.5	18.0	4.0
PTK	particles	2.2	7.3	16.9
VRB	verbs	16.5	19.8	6.4

Figure 1: Relative frequencies of the tags in the different corpora.

3.2 Compute speaker values

Apart from frequency and average length of words in a class, several other parameters were computed and averaged for every speaker (Figure 2). Important differences exist in the length of the turns, and also in the number of words per sentence. An average VM turn has more than 6 times as many words as an average TABA turn. While the length of a phrase is fairly equal between VM and CH, the number of phrases (and, thus, words) per sentence is higher for VM. Neither casual nor formal addressing are present in the TABA corpus (talking to a machine) while the VM setting (negotiation of business appointments) evokes formal speech. The CH dialogues are mostly between family members and close friends (Linguistic Data Consortium, 1997), and casual addressings are frequent. Variations in the number of common words can be related to the list of common words used, which was based on the VM corpus. The larger number of different words per speaker in the TABA corpus results from less words per speaker with less chance for repetition. Average word length and density are similar in all three corpora.

The correlation coefficients between parameters were computed for the three corpora. Those parameters that correlated well with another parameter (correlation coefficient > 0.6) were omitted from the pertinent corpus.

The correlation coefficients between WIP and WIS are strong for CH and moderate for VM, while those between PIS and WIS are moderate for CH and strong for VM. This indicates that longer sentences have longer phrases in the CH corpus but more phrases in the VM corpus. Different annotation styles and guidelines may have caused this phenomenon.

3.3 Principal component analysis

To normalize for the different ranges of the speaker-specific parameters the z scores (subtraction of the mean, division by standard deviation) are computed as input for the principal component analysis (PCA). The PCA was done with singular value decomposition on the data matrix. This is the preferred method for numerical accuracy

Par.	Meaning	VM	CH	TABA
SIT	sentences per turn (sentence end is marked by a question mark or colon in the VM and CH corpora, in the TABA corpus no sentence boundaries are labeled)	2.0	1.1	1.0
PIT	phrases per turn (in the TABA corpus no phrase boundaries are labeled)	4.3	1.9	1.0
WIT	words per turn	18.5	8.1	2.5
PIS	phrases per sentence	2.1	1.7	1.0
WIS	words per sentence	9.0	7.0	2.5
WIP	words per phrases	4.2	4.0	2.5
AWL	average word length 4.8	4.2	5.0	
CAS	casual addressing	0.0	13.0	0.0
FOR	formal addressing	7.0	5.0	0.0
DEN	ratio of “dense” words (Pieper, 1979) (attributive adjectives, nouns, and finite verbs)	0.3	0.3	0.3
DFW	ratio of different words	0.4	0.4	0.8
CWD	ratio of “common” words (words comprising 50 % of the VM corpus)	0.5	0.4	0.3

Figure 2: Average values of speaker-specific parameters for each corpus. Displayed are the values for the first quartile, the median, and the third quartile.

(Mardia et al., 1979).

The PCA results were used to assess the importance of the parameters. The important parameters (those that achieve high loads on the most important principal components) should not change much in dependence of the input set. Changes between the different corpora are likely regarding their structural differences, but ideally a stable set of parameters emerges that contains parameters important for all corpora.

The input partition was varied by changing the number of minimal words per speaker in order to check for the stability of the components and for the influence of the interaction length. Figure 3 displays some results. It can be seen that the important parameters for a corpus do not change much if the input set is varied. The principal components also show strong similarities within a corpus.

An interpretation of the components is always to be treated with caution. However, to ease further discussions, a tentative interpretation is given in Figure 4 for some of the principal components shown in Figure 3 where the interpretative labels can be motivated as follows:

adverbs: An adverb can take the position of a prepositional phrase or verb phrase, if an appropriate entity is present in the previous discourse context. Thus,

VM-1 100+ words, 816 speakers				
PC 1	PC 2	PC 3	PC 4	
CPRP 0.35	LVRB -0.35	WIT -0.35	CPRO -0.51	
CNOM 0.35	AWL -0.29	WIP -0.30	CART 0.35	
CADV -0.32	FOR -0.29	DFW 0.29	LNOM 0.32	
CNUM 0.30	CKON -0.28	FNUM -0.27	LART 0.31	
VM-2 400+ words, 410 speakers				
PC 1	PC 2	PC 3	PC 4	
AWL -0.33	LVRB -0.34	WIT 0.38	CPRO 0.41	
CPRP -0.31	CVRB -0.34	WIP 0.35	CKON -0.38	
CNOM -0.31	FOR -0.31	CWD 0.35	FPRO 0.29	
CITJ 0.30	LPRO -0.31	CPRO 0.29	CADV -0.28	
VM-3 800+ words, 148 speakers				
PC 1	PC 2	PC 3	PC 4	
AWL 0.32	CNUM -0.35	CPRO 0.40	CKON -0.44	
CPRP 0.29	FNUM -0.34	DFW -0.30	CADV -0.32	
CNOM 0.29	CVRB 0.32	WIP 0.29	CAS 0.29	
WIP 0.27	FVRB 0.31	WIT 0.27	WIT -0.29	
CH-1 100+ words, 159 speakers				
PC 1	PC 2	PC 3	PC 4	
WIT 0.35	CNOM 0.38	CADJ 0.36	CADV -0.44	
WIS 0.35	CPRO -0.34	FADJ 0.29	CVRB 0.39	
AWL 0.33	CART 0.31	LART 0.29	FADV -0.30	
CITJ -0.30	FPRO -0.27	FART -0.28	LART 0.29	
CH-2 400+ words, 150 speakers				
PC 1	PC 2	PC 3	PC 4	
WIS 0.35	CNOM 0.34	CVRB 0.41	LART -0.38	
WIT 0.35	CART 0.32	CADV -0.36	DFW -0.36	
AWL 0.33	CPRO -0.31	CPRO 0.33	FART 0.33	
CPTK -0.32	FART 0.30	CADJ -0.32	FADV 0.30	
CH-3 800+ words, 106 speakers				
PC 1	PC 2	PC 3	PC 4	
AWL -0.38	CNOM -0.41	CVRB 0.38	FART -0.41	
WIS -0.36	CPRP -0.30	CPRO 0.34	LART 0.39	
WIT -0.34	FADJ 0.29	CADJ -0.30	DFW 0.32	
CPTK 0.33	CADJ 0.29	FPRO 0.29	FADV -0.28	
TABA-1 30+ words, 780 speakers				
PC 1	PC 2	PC 3	PC 4	
CPRO -0.41	CWD -0.52	AWL 0.43	LITJ 0.42	
CVRB -0.38	CART -0.30	LADJ 0.40	CITJ 0.41	
WIT -0.36	AWL 0.29	CITJ -0.30	CADJ 0.35	
CNOM 0.28	CNOM 0.29	CPTK -0.30	LADJ 0.28	
TABA-2 40+ words, 409 speakers				
PC 1	PC 2	PC 3	PC 4	
CPRO -0.39	CWD 0.50	LITJ 0.53	LADJ 0.41	
WIT -0.35	AWL -0.39	CITJ 0.52	AWL 0.39	
CVRB -0.34	CPTK 0.32	CADJ 0.30	CART 0.33	
DFW -0.29	CNOM -0.30	LADJ 0.27	CITJ -0.31	
TABA-3 60+ words, 129 speakers				
PC 1	PC 2	PC 3	PC 4	
CPRO -0.38	AWL -0.50	CITJ 0.44	LADJ -0.39	
CVRB -0.35	CPTK 0.39	LITJ 0.42	CADJ -0.34	
WIT -0.35	CNOM -0.36	CART -0.34	CWD -0.30	
DFW -0.29	LADJ -0.35	CNUM 0.33	LNOM -0.30	

Figure 3: Component loads for the four most important principal components (PC) with varying input for the three corpora.

the number of adverbs/adjectives loads inversely to the number of prepositions and nouns (VM-1 PC 1, VM-3 PC 4, CH-1 PC 4, CH-2 PC 3, CH-3 PC 2).

pronouns: A pronoun or a noun can refer to a discourse entity, if that entity satisfies certain conditions. The number of pronouns loads inversely to the number of articles and nouns (VM-1 PC 4, CH-1 PC 2, CH-2 PC 2) and adverbs/adjectives (VM-2 PC 4, CH-3 PC 2). Pronominalization is only possible if the referred entity is mentioned in the very recent discourse context, ideally in the same turn; thus, longer turns favor pronominalization (VM-3 PC 3).

ellipses: Ellipses are incomplete sentences where redundant information is omitted (often verbs, VM-2 PC 2, VM-3 PC 2). Final articles, adverbs, and adjectives can indicate elliptic utterances (CH-1 PC 3, CH-2 PC 4, CH-3 PC 4).

turn complexity: Turns with pronouns and verbs are long and very likely contain a complete sentence (TABA-1 PC 1, TABA-2 PC 1, TABA-3 PC 1).

content words: The ratio of content words (less common words) is high (TABA-1 PC 2, TABA-2 PC 3).

Corpus	PC 1	PC 2	PC 3	PC 4
VM-1	adverbs	word length	turn length	pronouns
VM-2	word length	ellipses	turn length	pronouns
VM-3	word length	ellipses	pronouns	adverbs
CH-1	turn/word length	pronouns	ellipses	adverbs
CH-2	turn/word length	pronouns	adverbs	ellipses
CH-3	turn/word length	adverbs	pronouns	ellipses
TABA-1	turn complexity	content words	word length	interjections
TABA-2	turn complexity	content words	interjections	word length
TABA-3	turn complexity	word length	interjections	adjectives

Figure 4: Interpretation of the principal components displayed in Figure 3.

interjections: Interjections are rare in the TABA corpus. An interjection may, therefore, distinguish speakers (with interjections) from others (without interjections) (TABA-1 PC 4, TABA-2 PC 4, TABA-3 PC 3).

While some of these components are corpus-specific (e.g. *interjections*), others are important for all corpora (e.g. *word length* or *ellipses / turn complexity*) or, at least, for the syntactically complex human-human corpora (e.g. *pronouns* and *adverbs*).

The stability of the components was further checked by forming four subsets of speakers and applying the PCA to all 16 combinations of these subsets. A higher number of observations (speakers) results in higher stability. For VM-1 and TABA-1 with 800 observations, all four components appear in every subset. For VM-2 and TABA-2 with 400 observations, three common components exist while the fourth (least important component) varies. For CH-1 and VM-3 with approximately 160 speakers, two common components are found in all subsets.

3.4 Clustering

If a limited set of linguistically interpretable components exist, as has been argued for in the previous section, the question is whether speaker groups can be established, and whether unseen speakers can be reliably assigned the correct group.

To establish classes the k-means algorithm was employed. This algorithm works by repeatedly moving all cluster centers to the mean of their Voronoi sets. The algorithm stops, if no cluster center has changed during the last iteration or the maximum number of iterations is reached (Hartigan and Wong, 1979). The initial cluster centers are randomly assigned, thus, slightly different re-

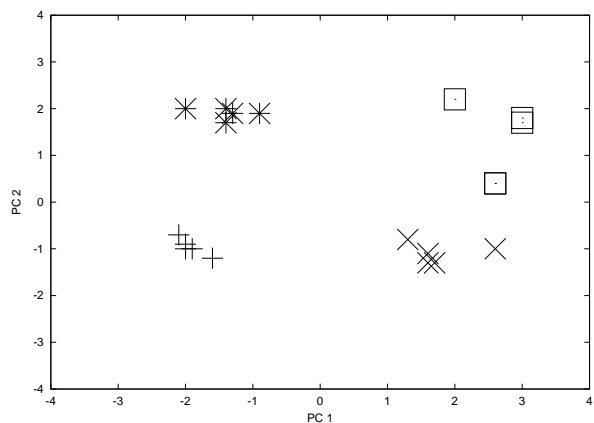


Figure 5: Cluster center distribution in the plane spanned by the first two components for five runs with varied input values.

sults are possible. The algorithm results in a predefined number of speaker clusters (Doux et al., 1997) that can be used to train automatic classifiers.

If a specific interpretation of the clusters for a given task is desired the clustering can be done by hand (i.e. by explicitly constructing borders between classes). In the data-driven approach taken here, however, such hand-crafted constraints were avoided (the same holds for the PCA which could also be replaced by explicit rules). Figure 5 shows a distribution of cluster centers for five different runs on the same data but with different initial cluster centers. The distribution displayed here in the plane of the first two components is fairly stable.

The four most important principal components (measured by their eigenvalue) computed for each speaker were used as input for the subsequent tests. This choice was motivated by the results of the component stability experiments described above. The current set of speakers does not support a more fine-grained distinction. Furthermore, it is unlikely that more dimensions will be useful for applications.

3.5 Classification

A correct classification of individual linguistic styles described by the parameter set used in this experiment means that a speaker is put into the same class by the cluster analysis and the automatic classification. To test this hypothesis the turns for each speaker were alternately divided into two sets of the same size. The first set was used for the training of the classifier (calculate PCA, estimate clusters to obtain classes). The second set served as a test set. If the error rate (different classifications for the training set and the test set) is substantially lower than chance one can state that the parameters can be used to reliably discriminate between speaker classes.

Neural networks were used for automatic classification. For training, two sets of input vectors were generated from the original training set. Every n -th pattern

became part of a development set (n is 5 for all experiments described here). The output of the nets consisted of m output values (one for each class which is either 0 or 1). Fully connected feed forward nets with standard back propagation were trained until the error averaged over the last three runs on the development set begins to increase (overtraining). The net topology had one input layer, one output layer, and one hidden layer with the same number of nodes as the input layer.

The speaker specific values for the test set and the values for the principal components were computed and used as input for the neural network. The classification of the network was judged correct for one speaker, if the cluster determined by the cluster analysis on the training set was equal to the class predicted by the neural network on the test set.

The results are displayed in Figures 6 and 7. While the results for the TABA corpus are only slightly above chance level (25 %), the results for the VM and CH corpora indicate that for human-human corpora a speaker can be fairly reliably classified. If not enough turns or words are available the result decreases. The result decreases also if the number of speakers is too small (below 30, results are not shown here).

4 Discussion

These results indicate that style classification in a spoken dialogue system is only feasible if the number of interactions (turns) is sufficiently high and linguistically rich. If this is the case, however, speakers can be classified according to simple part-of-speech distribution and turn length parameters which can be computed automatically (Section 3.1), which can be automatically combined to components interpretable as linguistic style indicators (Section 3.3), that can be employed to automatically group speakers into classes (Section 3.4), which, in turn, can be used by automatic learning methods to classify unseen turns from one speaker into the same class as the reference turns from the same speaker used during the training (Section 3.5).

5 Applications

Several possible applications for linguistic style information exist in a spoken dialogue system. Among these are

- style-specific language models,
- style-specific grammars,
- input to a general user model (certain elements of linguistic style can be related to paradigm or task knowledge, e.g. turn length, number of content words),
- influencing the style of a language generation module.

An exploratory experiment was undertaken for the first application.

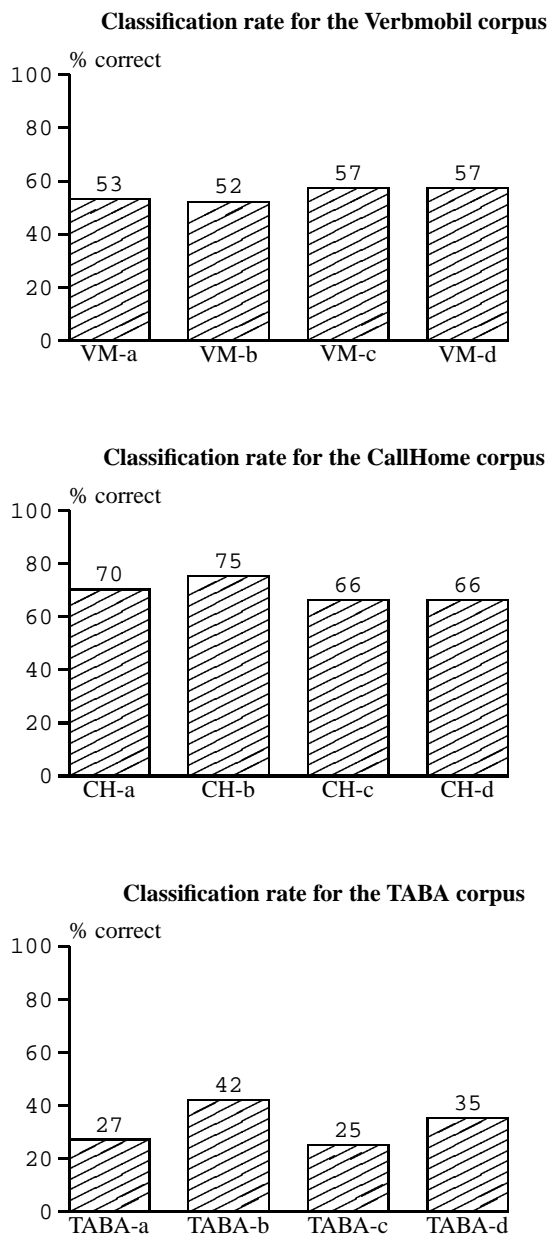


Figure 6: Rate of correct classifications for the different corpora, displayed for varying input sets (see Figure 7 for a description).

5.1 Material and Method

The three corpora described above were used for the investigation. For the CH and VM corpora all speakers with more than 100 words were used; this threshold was set to 40 for the TABA corpus. The clustering method described above based on the results of a PCA was employed to group the speakers into 2 to 4 clusters based on 2 to 4 factors. The k-means clustering algorithm with a fixed set of clusters is initialized by random cluster centers. Thus, two subsequent runs give slightly different results. Therefore, each run was performed twice.

The four corpora plus their combination were divided

Task	min. Words	min. Turns	Speakers	Class.
VM-a	100	20	188	53
VM-b	100	30	109	52
VM-c	400	20	125	57
VM-d	400	30	87	57
CH-a	100	40	144	70
CH-b	100	60	60	75
CH-c	400	40	99	66
CH-d	400	60	47	66
TABA-a	30	6	158	27
TABA-b	30	8	65	42
TABA-c	40	6	123	25
TABA-d	40	8	53	34

Figure 7: Data sets used in the classification task with minimal number of words and turns per single speaker as constraints, and the resulting speaker number and classification rate.

into 10 sets for a ten-fold cross-validation. One set served as development set for parameter optimization, another set was the test set, and the remaining eight sets were used to train the language models. In each run five standard bigram language models were trained, one for each speaker-specific corpus and one for their combination. The perplexity was calculated for the pertinent test sets.

In a subsequent step for each speaker-specific corpus the general and the specific language models were linearly interpolated; the interpolation factor was iterated over the values 0, 0.2, 0.4, 0.6, 0.8, and 1.0. The interpolation factor which gave the best results (smallest perplexity) for the development set was taken to compute the perplexity on the test set.

5.2 Results

When interpreting the results one has to keep a few things in mind:

- The assumption that the correct style class of a speaker is known in advance is not likely to be true in real systems. A few turns have to be analyzed in order to perform a reasonable classification.
- The parameters used for classification (distributions of part-of-speech items, length parameters etc.) are only very loosely related to the probability of word sequences.
- The classes are not optimized to yield maximal gain in perplexity.
- The language models are rather simple.

The global results are displayed in Figure 8. The general model (with 4 times the training material than the special models) gives better results, except for the TABA corpus which is probably sufficiently constrained and simply structured to make up for the decrease in training material. This is in line with results described in Klarner (1997). The interpolated model has a significantly lower perplexity than the general model alone, but the gain is

Corpus	General	Special	Sig.	Interpol.	Sig.
CallHome	215.1	236.4	yes	210.3	yes
Verbmobil	106.2	115.0	yes	103.5	yes
TABA	26.1	24.8	yes	23.1	yes

Figure 8: Mean perplexity values for all comparison runs ($n=324$) between style-specific and general language models (top) and between interpolated and general language models (bottom). Significance was calculated by the paired t-test (one-sided, $\alpha < 0.05$, $df = 35$).

so small that it is unlikely to improve recognition results. With all the caveats listed above one can conclude that determination of linguistic style in the way described in this document does not dramatically improve recognition results.

6 Conclusion

This investigation showed that

- numerical parameter values can be computed for speakers in spoken dialogue corpora,
- these parameters can be reduced to linguistically interpretable factors by means of a PCA,
- stable classes can be constructed from these factors by cluster analysis,
- unseen class members can be reliably classified by trained neural networks if the data is linguistically rich,
- style-specific language models reduce perplexity only marginally.

This process has been applied to three different corpora. It has been shown that it works in principle. Further improvements may be obtained by optimizing the procedure according to specific needs (e.g. very quick classification, recognizing a speaker from a small set of possible speakers) which depend on the application.

The methods can not only be applied to classify speakers according to their style, but also to recognize text genre or speech act types.

7 Acknowledgment

This research was conducted within the SmartKom project and partly funded by the German ministry of Research and Technology.

References

- Harald Aust, Martin Oerder, Frank Seide, and Volker Steinbiss. 1995. The Philips automatic train timetable information system. *Speech Communication*, 17:249–262.
- Susan E. Brennan. 1996. Lexical entrainment in spontaneous dialog. In *Proc. ISSD 96*, pages 41–44.
- Penelope Brown and Stephen C. Levinson. 1987. *Politeness: some universals in language usage*. Cambridge University Press, Cambridge.

- Christine Doran, John Aberdeen, Laurie Damianos, and Lynette Hirschman. 2001. Comparing several aspects of human-computer and human-human dialogues. In *Proc. Second SIGDial Workshop on Discourse and Dialogue*, Aalborg.
- Anne-Claude Doux, Jean-Philippe Laurent, and Jean-Pierre Nadal. 1997. Symbolic data analysis with the k-means algorithm for user profiling. In *User Modeling: Proceedings of the UM 97*, pages 359–361.
- Laurel Fais and Kyung Ho Loken-kim. 1995. Lexical accommodation in human-interpreted and machine-interpreted dual-language interactions. In *Proc. ESCA Workshop on Spoken Dialogue Systems*, page 69, Vigsø, Denmark.
- Laurel Fais, Kyung-Ho Loken-Kim, and Tsuyoshi Morimoto. 1996. How Many Words is a Picture Really Worth? In *Proc. ICSLP'96*, Philadelphia, USA.
- Joakim Gustafson, Anette Larsson, Rolf Carlson, and K. Hellman. 1997. How do system questions influence lexical choices in user answers? In *Proc. Eurospeech '97*, pages 2275–2278, Rhodes, Greece, September.
- J. A. Hartigan and M. A. Wong. 1979. A k-means clustering algorithm. *Applied Statistics*, 28:100–108.
- Jussi Karlgren and Douglas Cutting. 1994. Recognizing text genres with simple metrics using discriminant analysis. In *Proc. Coling 94*, Kyoto.
- Jussi Karlgren. 1994. Stylistic variation in an information retrieval experiment. In *Proceedings of NEMLAP-2*.
- Martin Klarner. 1997. Klassifikation von Sprechstilen mit linguistischem Wissen. Master's thesis, IMMD, Friedrich-Alexander-Universität, Erlangen.
- Jill Fain Lehman and Jaime G. Carbonell. 1989. Learning the user's language: A step towards automated creation of user models. In Alfred Kobsa and Wolfgang Wahlster, editors, *User Models in Dialog Systems*. Springer Verlag, Berlin—New York.
- Linguistic Data Consortium. 1997. Callhome transcript corpus of German telephone speech.
- K. V. Mardia, J. T. Kent, and J. M. Bibby. 1979. *Multivariate Analysis*. Academic Press, London.
- T. C. Mendenhall. 1887. The characteristic curves of composition. *Science*, 9(214):237–249.
- Michio Okada, Noriko Suzuki, and Masaaki Date. 1999. Social bonding in talking with social autonomous creatures. In *Proc. Eurospeech 99*, volume 4, pages 1731–1734, Budapest.
- Sharon L. Oviatt and Philip R. Cohen. 1991. Discourse structure and performance efficiency in interactive and non-interactive spoken modalities. *Computer Speech and Language*, 5:297–326.
- Sharon L. Oviatt, Philip R. Cohen, and Micelle Wang. 1994. Toward interface design for human language technology: Modality and structure as determinants of linguistic complexity. *Speech Communication*, 15:283–300.
- Daniel Paiva. 2000. Investigating style in a corpus of pharmaceutical leaflets: Results of a factor analysis. In *Proc. Annual Meeting of the ACL, Student Session*, pages 52–59, Hong Kong.
- Ursula Pieper. 1979. *Über die Aussagekraft statistischer Methoden für die linguistische Stilanalyse*. Narr, Tübingen.
- Elaine M. Rich. 1979. User modeling via stereotypes. *Cognitive Science*, 3:329–354.
- Klaus Ries. 1999. Towards the detection and description of textual meaning indicators in spontaneous conversations. In *Proc. Eurospeech 99*, volume 3, pages 1415–1418, Budapest.
- Anne Schiller, Simone Teufel, Christine Thielen, and Christine Stöckert. 1995. Vorläufige guidelines für das taggen deutscher textcorpora mit stts. Technical report, IMS, Universität Stuttgart.
- Helmut Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. In *International Conference on New Methods in Language Processing*, Manchester, UK, September.
- W. Wahlster, N. Reithinger, and A. Blocher. 2001. Smartkom: Multimodal communication with a life-like character. In *Proceedings of the Eurospeech 2001*, pages 1547–1550, Aalborg, Denmark.
- Wolfgang Wahlster. 1993. VERBMOBIL—translation of face-to-face dialogs. In Otthein Herzog, Thomas Christaller, and Dieter Schütt, editors, *Grundlagen und Anwendungen der Künstlichen Intelligenz. 17. Fachtagung für Künstliche Intelligenz*, pages 393–402, Berlin, Heidelberg, New York. Springer. Informatik Aktuell.
- Marilyn A. Walker, Janet E. Cahn, and Stephen J. Whittaker. 1997. Improvising linguistic style: Social and affective bases of agent personality. In W. Lewis Johnson and Barbara Hayes-Roth, editors, *Proceedings of the 1st International Conference on Autonomous Agents*, pages 96–105, New York, February 5–8. ACM Press.
- Maria Wolters and Mathias Kirsten. 1999. Exploring the use of linguistic features in domain and genre classification. In *Proc. EACL 99*, Bergen.